Trade-Offs Between Approximation and Generalization in Learning Systems

Shahin Shahrampour, Department of Industrial & System Engineering
(Joint work with Yinsong Wang)

T3: TEXAS A&M TRIADS FOR TRANSFORMATION
A President's Excellence Fund Initiative

Random Features and Kernel Approximation

Motivation: Random features has been widely used for kernel approximation in large-scale machine learning. A number of recent studies have explored *data-dependent* sampling of features, modifying the stochastic oracle from which random features are sampled. While proposed techniques in this realm improve the approximation, their application is limited to a specific learning task. In this work, we propose a general scoring rule for sampling random features, which can be employed for various applications with some adjustments.

Notation:
$\mathrm{tr}[\cdot]$ denotes the trace operator. $\mathbb{E}[\cdot]$ denotes the expectation operator. $[\mathbf{A}]_{ij}$ denotes the ij-th entry of matrix $\mathbf{A}$. $\mathbf{\Sigma}_{xy}$ denotes the covariance matrix of random variables $\mathbf{X}$ and $\mathbf{Y}$.

Random features and kernel approximation:

- $\{\mathbf{x}_i\}_{i=1}^n$ is a set of given points where $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ for any $i \in \{1, \ldots, n\}$.
- Consider any kernel function in the following form

$$k(\mathbf{x}, \mathbf{x}') = \int_\Omega \phi(\mathbf{x}, \omega)\phi(\mathbf{x}', \omega)p(\omega)d\omega, \tag{1}$$

where $\phi(\mathbf{x}, \omega) : \mathbb{R}^{d_x} \to \mathbb{R}$ is a feature map parameterized by $\omega \in \mathbb{R}^{d_x}$. Following (1), the kernel function can be approximated as

$$k(\mathbf{x}, \mathbf{x}') \approx \frac{1}{M}\sum_{m=1}^{M} \phi(\mathbf{x}, \omega_m)\phi(\mathbf{x}', \omega_m),$$

$\{\omega_m\}_{m=1}^M$ are independent samples from $p(\omega)$, called *random features*.

- Let us now define

$$\mathbf{z}(\omega) \triangleq [\phi(\mathbf{x}_1, \omega), \ldots, \phi(\mathbf{x}_n, \omega)]^\top.$$

Then, the kernel matrix $[\mathbf{K}]_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ can be approximated with $\mathbf{Z}\mathbf{Z}^\top$ where $\mathbf{Z} \in \mathbb{R}^{n\times M}$ is defined as

$$\mathbf{Z} \triangleq \frac{1}{\sqrt{M}}[\mathbf{z}(\omega_1), \ldots, \mathbf{z}(\omega_M)]. \tag{2}$$

- The low-rank approximation above can save significant computational cost when $M \ll n$.

The General Scoring Rule for random features

Let $\mathbf{B}$ be a positive definite matrix. We propose the following score function for any $\omega \in \Omega$

$$q(\omega) \triangleq \frac{p(\omega)\mathbf{z}^\top(\omega)\mathbf{B}\,\mathbf{z}(\omega)}{\mathbb{E}_{p(\omega)}[\mathbf{z}^\top(\omega)\mathbf{B}\,\mathbf{z}(\omega)]} = \frac{p(\omega)\mathbf{z}^\top(\omega)\mathbf{B}\,\mathbf{z}(\omega)}{\mathrm{tr}[\mathbf{K}\mathbf{B}]}, \tag{3}$$

where $p(\omega)$ is the original probability density of random features.

- The key advantage of the score function is that $\mathbf{B}$ can be designed to improve sampling depending on the learning task.
- Setting $\mathbf{B} = (\mathbf{K} + \lambda\mathbf{I})^{-1}$ in (3) can precisely recover leverage score (LS) sampling

$$q_{LS}(\omega) = \frac{p(\omega)\mathbf{z}^\top(\omega)(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{z}(\omega)}{\mathrm{tr}[\mathbf{K}(\mathbf{K} + \lambda\mathbf{I})^{-1}]}. \tag{4}$$

- Setting $\mathbf{B} = \mathbf{y}\mathbf{y}^\top$ in (3) can equivalently recover Energy-based Exploration of Random Features (EERF)

$$q_{\mathrm{EERF}}(\omega) \propto \left|\frac{1}{n}\sum_{i=1}^{n} y_i\phi(\mathbf{x}_i, \omega)\right|. \tag{5}$$

- If random features are sampled from the score function (3), the transformed matrix will be

$$\widetilde{\mathbf{Z}} \triangleq \frac{1}{\sqrt{M}}\left[\sqrt{\frac{p(\omega_1)}{q(\omega_1)}}\mathbf{z}(\omega_1), \ldots, \sqrt{\frac{p(\omega_M)}{q(\omega_M)}}\mathbf{z}(\omega_M)\right],$$

to form an unbiased approximation of the kernel matrix $\mathbf{K}$.

Adaptation to Canonical Correlation Analysis

Formulation of Canonical Correlation Analysis

- Linear Canonical Correlation Analysis is a method of correlating linear relationships between two multi-dimensional random variables $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top \in \mathbb{R}^{n\times d_x}$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^\top \in \mathbb{R}^{n\times d_y}$. The canonical correlations are the eigenvalues of the following matrix

$$\begin{bmatrix} (\mathbf{\Sigma}_{xx} + \mu_x\mathbf{I})^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{\Sigma}_{yy} + \mu_y\mathbf{I})^{-1} \end{bmatrix}\begin{bmatrix} \mathbf{0} & \mathbf{\Sigma}_{xy} \\ \mathbf{\Sigma}_{yx} & \mathbf{0} \end{bmatrix}. \tag{6}$$

- Kernel Canonical Correlation Analysis (KCCA) correlates nonlinear relationships of two random variables in Reproducing Kernel Hilbert Space. The kernel canonical correlations are the eigenvalues of the following matrix

$$\begin{bmatrix} (\mathbf{K}_x + \mu_x\mathbf{I})^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{K}_y + \mu_y\mathbf{I})^{-1} \end{bmatrix}\begin{bmatrix} \mathbf{0} & \mathbf{K}_y \\ \mathbf{K}_x & \mathbf{0} \end{bmatrix}.$$

- We assume $\mu_x = \mu_y = \mu$.
- Maximizing the total canonical correlations will then be equivalent to maximizing

$$\mathrm{tr}[(\mathbf{K}_x + \mu\mathbf{I})^{-1}\mathbf{K}_y(\mathbf{K}_y + \mu\mathbf{I})^{-1}\mathbf{K}_x].$$

Proposition: To maximize total canonical correlations, we should use the following center matrix

$$\mathbf{B} = (\mathbf{K}_x + \mu\mathbf{I})^{-1}\mathbf{K}_y(\mathbf{K}_y + \mu\mathbf{I})^{-1}. \tag{7}$$

Algorithm

Algorithm 1 Optimal Randomized Canonical Correlation Analysis 2 (ORCCA2)

Input: $\mathbf{X} \in \mathbb{R}^{n\times d_x}, \mathbf{Y} \in \mathbb{R}^{n\times d_y}$, the feature map $\phi(\cdot,\cdot)$, an integer M_0, an integer M, the prior densities $p_x(\boldsymbol{\omega})$ and $p_y(\boldsymbol{\omega})$, the parameter $\mu > 0$.

1: Draw samples $\{\boldsymbol{\omega}_{x,m}\}_{m=1}^{M_0}$ according to $p_x(\boldsymbol{\omega})$, and $\{\boldsymbol{\omega}_{y,m}\}_{m=1}^{M_0}$ according to $p_y(\boldsymbol{\omega})$, respectively.

2: Construct the matrices

$$\mathbf{Q} = (\mathbf{Z}_x^\top\mathbf{Z}_x + \mu\mathbf{I})^{-1}\mathbf{Z}_x^\top\mathbf{Z}_y$$
$$\mathbf{P} = (\mathbf{Z}_y^\top\mathbf{Z}_y + \mu\mathbf{I})^{-1}\mathbf{Z}_y^\top\mathbf{Z}_x.$$

where $\mathbf{Z}_x$ and $\mathbf{Z}_y$ are defined in (1).

3: Let for $i \in [M_0]$

$$\widehat{q}_x(\boldsymbol{\omega}_{x,i}) = \frac{[\mathbf{QP}]_{ii}}{\mathrm{Tr}\,[\mathbf{QP}]}.$$

The new weights $\widehat{\mathbf{q}}_x = [\widehat{q}_x(\boldsymbol{\omega}_1), \ldots, \widehat{q}_x(\boldsymbol{\omega}_{M_0})]^\top$.

4: Let for $i \in [M_0]$

$$\widehat{q}_y(\boldsymbol{\omega}_{y,i}) = \frac{[\mathbf{PQ}]_{ii}}{\mathrm{Tr}\,[\mathbf{PQ}]}.$$

The new weights $\widehat{\mathbf{q}}_y = [\widehat{q}_y(\boldsymbol{\omega}_1), \ldots, \widehat{q}_y(\boldsymbol{\omega}_{M_0})]^\top$.

5: Select top M features with the highest scores from each of the pools $\{\boldsymbol{\omega}_{x,i}\}_{i=1}^{M_0}$ and $\{\boldsymbol{\omega}_{y,i}\}_{i=1}^{M_0}$, according to the new scores $\widehat{\mathbf{q}}_x$ and $\widehat{\mathbf{q}}_y$ to construct the transformed matrices $\widehat{\mathbf{Z}}_x \in \mathbb{R}^{n\times M}$ and $\widehat{\mathbf{Z}}_y \in \mathbb{R}^{n\times M}$, respectively, as in (1).

Output: Linear canonical correlations between $\widehat{\mathbf{Z}}_x$ and $\widehat{\mathbf{Z}}_y$ (with regularization parameter μ) as in (6).

- If we use linear kernel for Y domain, the center matrix can be simplified and we call that algorithm ORCCA1.
- ORCCA2 is designed for nonlinear kernel in $\mathbf{Y}$ domain.
- The algorithms are derived through replacing the true kernel $\mathbf{K}$ in (7) with $\mathbf{Z}\mathbf{Z}^\top$.

Numerical Experiments

Benchmark Algorithms:

- **Random Fourier Features (RFF)** with $\phi = \cos(\mathbf{x}^\top\omega + b)$ as the feature map to approximate the Gaussian kernel.
- **Orthogonal Random Features (ORF)** with $\phi = [\cos(\mathbf{x}^\top\omega), \sin(\mathbf{x}^\top\omega)]$ as the feature map.
- **Leverage Score (LS)** with $\phi = \cos(\mathbf{x}^\top\omega + b)$ as the feature map.
- **Energy-based Exploration of Random Features (EERF)** with $\phi = \cos(\mathbf{x}^\top\omega + b)$ as the feature map. EERF only works for supervised learning and is only suitable for comparison with ORCCA1.

Practical Consideration:

- We work with empirical copula transformation of datasets.
- For $\mathbf{X}$ domain, the variance of random features σ_x is set to be the inverse of mean-distance of 50-th nearest neighbour (in Euclidean distance), $\sigma_y = \sigma_x$.
- For EERF, LS, and ORCCA2, the pool size is $M_0 = 10M$.
- The regularization parameter λ for LS is chosen through grid search.
- The regularization parameter is set to by $\mu = 10^{-6}$.

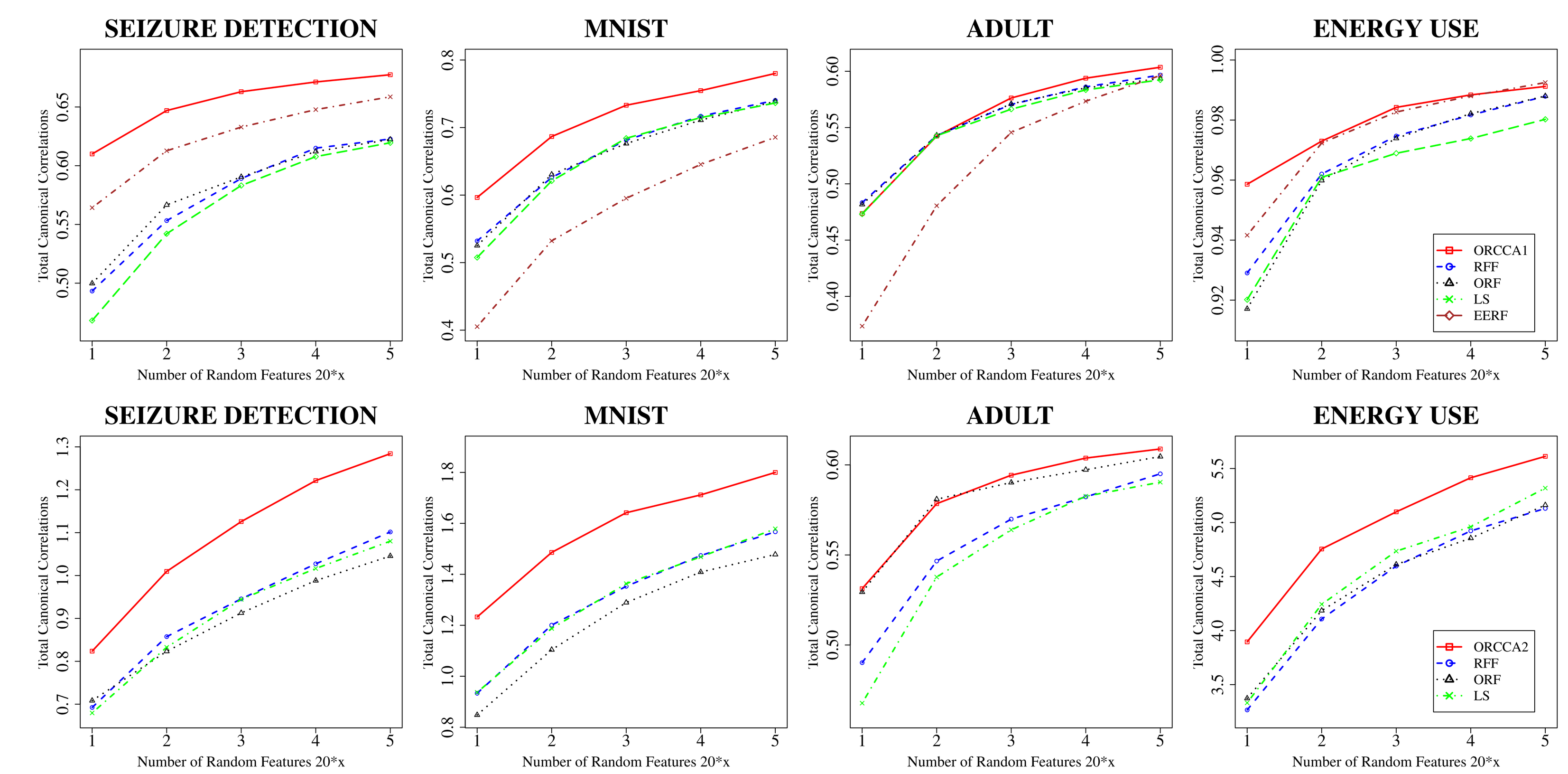

Figure: The plot of total canonical correlations obtained by different algorithms versus the number of features.

Performance: Judging from the plots, choosing $\mathbf{B}$ according to our theory will give a significant boost in increasing the total canonical correlations.

Current and Future Directions

Together with another member of the triad, **Dr. Simon Foucart** (MATH), we are investigating learning problems from an Optimal Recovery perspective. This framework considers learning with non-random data, where generalization in statistical sense is no longer applicable. We have looked at the notion of worst-case error in Hilbert spaces and showed that Optimal Recovery provides a formula which is user-friendly from an algorithmic point-of-view. Our future directions include specific problems arising in Optimal Recovery, such as robustness to measurement noise, over-parameterized learning, nonlinear hypothesis classes, and beyond.

Acknowledgement: We gratefully acknowledge the support of Texas A&M Triads for Transformation (T3) program. This funding has supported two graduate students **Yinsong Wang** (ISEN) and **Chunyang Liao** (MATH) towards their Ph.D. degrees.